Testing for Independence

1. Does ice cream preference affect performance on Math 10B midterms? A GSI surveys their 50 students on whether they like vanilla or chocolate more. Out of the students who did well, 15 prefer vanilla and 10 prefer chocolate; out of the remaining students, 11 prefer vanilla and the rest prefer chocolate.

   a) What are the hypotheses you are testing?

   b) Which test should you use here and why?

   c) What is the *p-value* of your experiment?

   d) What should you conclude?

2. Darwin wonders if there is any relationship between the colors of sparrows he observes (white, brown, black, or red) and the shape of their beak (flat or curved). He records the features of 100 sparrows in the following chart. What conclusions should he draw about his hypothesis? (Hint: You will need to fix the shape of the table.)

Red Flat	Red Curved	Brown Flat	Brown Curved	Black Flat	Black Curved	White Flat	White Curved
10	15	13	7	11	19	6	19

Correlation & Best Fit Lines

3. Suppose you want to know whether performance on Quiz 1 is correlated with performance on Quiz 12. You randomly choose 5 students' quiz scores and get the following values.

Student	Quiz 1	Quiz 12
A	7	9
B	12	11
C	6	5
D	11	10
E	4	5

   a) Make a (rough) scatter plot of these scores. Do they seem to have a strong linear relationship? What would you guess the *correlation coefficient* r is approximately?

   b) Calculate the sample standard deviation for Quiz 1 (σ_x). Do the same for Quiz 12 (σ_y).

   c) Calculate the *sample covariance* for the two variables.

   d) What is the correlation coefficient equal to? What does this tell us about the data?

4. Using the same data as above, we now want to make the best prediction for how someone who got an 8 on Quiz 1 would've scored on Quiz 12. Do this by first finding the *line of best fit* ($y = ax + b$) using the basic formulas for the MLE of a and b (use what you calculated above).

5. You are interested in whether two variables (x and y) are correlated, but due to budget constraints, you can only collect two data points. You get the data points $(1, 8)$ and $(3, 4)$.

   a) What do you guess the correlation coefficient will be?

   b) Calculate r and explain what this means about your data.

   c) Find the line of best fit.

6. Is there a relationship between the amount of antibody A and antibody B in a sick patient? You take antibody A and B counts per milliliter from 4 patients (in reality you will have a much, much larger sample size).

Patient	Antibody A	Antibody B
A	120	100
B	95	110
C	110	130
D	105	80

   a) What do you guess the correlation coefficient will be?

   b) What is the correlation coefficient?

   c) What will our line of best fit look like? If someone has an Antibody A count of 100, should we feel confident in guessing their Antibody B count?